A hybrid combinatorial method for docking a single-stranded RNA in a protein pocket at the thermodynamic equilibrium

Chinmay Singhal [1], Yann Ponty [2,3], Isaure Chauvot De Beauchêne * [4]

[1] AMIBio team, INRIA Saclay – INRIA – France
[2] Laboratoire d'informatique de l'école polytechnique [Palaiseau] (LIX) – CNRS : UMR7161, Polytechnique - X – Route de Saclay 91128 PALAISEAU CEDEX, France
[3] AMIB (INRIA Saclay - Ile de France) – Université Paris XI - Paris Sud, CNRS : UMR8623, Polytechnique - X, INRIA – Bât. Alan Turing ; Campus de l'Ecole Polytechnique ; 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[4] LORIA – CNRS : UMR7503 – France

INTRODUCTION

Protein-RNA complexes participate in many aspects of cell regulation, and their atomistic structural description is crucial to understand, modulate or engineer the recognition mechanism. As the experimental resolution of their structure is arduous, computational protein-RNA docking methods have been developed, that aim at modeling a 3D assembly by assembling structures of each isolated constituent. Yet for highly flexible objects like single-stranded RNA (ssRNA), the isolated structure of the whole molecule can adopt an ensemble of conformations too large to be experimentaly solved or computationaly modeled. Therefore, if somehow successful on structured RNAs, classical computational docking methods cannot handle the flexibility of ssRNA.

We have recently proposed an original fragment-based approach to accurately model ssRNA-protein complexes from protein structure and RNA sequence, consisting in (i) cutting the RNA sequence into trinucleotides overlapping by 2 nucleotides, represented by a fragment library built from known protein-RNA structures; (ii) docking each conformer-ensemble separately onto the protein; (iii) assembling the spatially compatible poses into an RNA chain.

This method can either blindly determine the RNA-binding site with high specificity [de_Beauchene, de_Vries, Zacharias, PloS 2016], or model the full RNA based on protein-RNA contacts predicted by homology [de_Beauchene, de_Vries, Zacharias, NAR 2016]. In the absence of known contacts, the number of compatible fragment chains is beyond the reach of brute force approaches.

Here, we present an improved method capable of modeling the full bound ssRNA without homology information. Improvements include:

i. a new docking protocol for sampling deep binding pockets;

ii. a stochastic backtracking algorithm for unbiased sampling of chains from the fragment connectivity graph, after computing the partition function of each pose by dynamic programming;

*Speaker

iii. a combination of filters based on biophysical characteristics of the binding site.

As a proof-of-principle, we successfully applied this method on a poly-U ssRNA inserted in the deep cavity of an exonuclease. The accuracy of 4 Å RMSD reached for this 10-mer ssRNA is far beyond the reach of any other docking program.

RESULTS

To evaluate the quality of our docking results, we computed the RMSD (Root Mean Squared Deviation) of the fragment poses or the total RNA chains to their reference position in the crystallographic structure. Given the very high flexibility and the hybrid size of our ssRNA , we adapted the classical acceptance criteria of 2 and 10 Å used for macromolecules and small-ligand docking, toward 3 and 5 Å for fragments and 8-mer chains respectively.

1. A new docking protocol for buried binding

For the fragment assembly to succeed, the sampling of each individual fragment from the sequence needs to be sufficiently comprehensive. We formerly used the ATTRACT docking software [Zacharias, Proteins 2003], which performs a minimization of the intermolecular energy in an empirical force field, starting from random positions of a ligand around a receptor. As the ligand can quickly be trapped in local minima of the energy landscape at a protein surface, the number of starting positions must be large to increase the probability of the ligand to find the global minimum. Our previous ATTRACT protocol for docking RNA on a globular protein surface started from 3.10^7 positions, among which the 2445 UUU conformers of our fragment library were randomly distributed. The 10^6 best-scored poses were retrieved. Yet for docking inside the deep buried cavity of our exonuclease, this sampling was not large enough. Only a small fraction of the starting positions could enter the cavity without getting stuck at the protein surface, resulting in only $0 - 42$ correct poses per fragment, with a best-RMSD up to 3.4 Å.

Therefore, we developed DeepATTRACT, a new protocol for docking inside deep cavities, and compared the sampling quality with the previous ATTRACT protocol. DeepATTRACT uses the detection of pocket points by the POCASA server [Yu, Zhou, Tanaka, Yao, Bioinformatics 2010] and selects as starting positions the points with enough neighors to accommodate a trinucleotide. The number of such points (5682) being too large for all the 2445 UUU conformers to be tested at each point, we used a "hierarchical sampling":

i. The library conformers were clustered by RMSD in 108 clusters;

ii. Each cluster center was placed at each starting point with 32 different orientations;

iii. Each combination (point * conformers * orientations) was scored with the ATTRACT function;

iv. When a good score (low energy) was found, each conformer in the same cluster was placed at the same position;

v. The new combinations were shortly minimized and re-scored, and the 10^6 best-scored poses were retrieved

With this new protocol, $54 - 644$ acceptable poses were found for each fragment, with a best-

RMSD in range 1.0 – 2.6 Å. As expected, this strong overall improvement over the previous ATTRACT protocol is particularly pronounced for the four most buried fragments: the number of acceptable poses inproved from 0 - 2 to 56 – 644, and the best-RMSD from 2.1 – 3.4 Å to 1.2 – 2.6 Å.

2. Assembly by stochastic backtracking

We then searched chains of compatible docking poses with a low total binding energy. As a first approximation, we used the ATTRACT score as a proxy for the pose binding energy, and additively defined over assemblies. The compatibility criteria between two successive poses was defined as an RMSD of the two overlapping nucleotides below 2 Å. Assembling 10^6 poses per fragment would lead to 10^48 possible 8-fragments chains. To retrieve the most probable assemblies while avoiding a brute-force enumeration, we used a new algorithm for unbiased sampling of chains.

The connectivity of each pair of poses was evaluated, resulting in a directed graph of connected poses. The partition function Z was computed over the set of all admissible assemblies, using dynamic programming. As a side product, the algorithm computes the exact Boltzmann probability of a given pose to participate in a downstream path (cf Methods). It can then be adapted to perform a stochastic sampling of the Boltzmann ensemble, resulting in a good approximation of the Boltzmann ensemble of low-energy. We iterated our sampling procedure and obtained 10^5 chains.

By repeating this sampling, we obtained 10^5 chains. After averaging the coordinates of overlapping nucleotides, we obtained a best RNA at 2.2 Å RMSD from the reference structure. But the fraction of acceptable models was low (3), and the scoring function of ATTRACT is not precise enough to select the best models, requiring the use of more effective filters.

3. Enrichment by combinable filters

To enrich the fraction of correct models, we used general and system-specific knowledge to define geometric constraints as filters:

(a) Mg2+ ions are well-known for being chelated by RNA phosphate groups and to stabilize RNA-protein complexes. One such ion is present at the bottom of the exonuclease pocket. We imposed the last RNA phosphate of our chain to be within 7 Å from it.

(b) Aromatic rings at protein-RNA interfaces are well-known for establishing stacking interactions with RNA bases. One aromatic ring is present at the entrance of the exonuclease binding pocket of our exonuclease. Based on the pocket size and the average nucleotide size, we imposed the 1st base of our 10-mer RNA chain to be within 5 Å from the aromatic ring.

(c) We assume a linear ssRNA , i.e. nucleotides from distinct fragment at more than 6 Å from each other.

Applying each filter separately retained pools with 0.5 – 12% correct models (enrichment x1.6 – x34). Combining two filters retained pools with 12 – 50% correct models (x2.7 – x163). The most effective filters were (b), (a) then (c), (c) being mosty redundant with (b). Finally, combining the three filters retained only one model, with an RMSD of 4.0 Å.

The complete process took few CPU hours.

DISCUSSION AND PERSPECTIVES

We present a method capable of modeling a protein-bound ssRNA based on the protein structure and ssRNA sequence. First, our new docking protocol for docking RNA fragments inside deep pockets improved the sampling quality for all buried fragments. Second, a new stochastic backtracking algorithm to perform unbiased sampling from a connectivity graph of the docked fragments generated near-native RNA chains among 100,000 samples. Finally, an efficient and effective filtering procedure to incorporate knowledge on the protein-ssRNA system led to a fraction in correct models of up to 50–100% after applying 2-3 filters. As a first proof-of-principle, the method could model *ab initio* a 10-mer bound ssRNA, an unprecedented length far beyond the reach of standard small-molecule or macromolecular docking programs.

However, several limitations must be regarded:

First, filters such as those used on that particular case are not always available solely from the knowledge of the protein structure. Additional experimental data (mutagenesis, cross-linking...) can be required. The advantage of this sample-then-filter approach compared to a constrained sampling is to be able to predict, from the initial sample, which set of experiments would best partition it. This reduces the number of experiments required for a given targeted enrichment factor.

Second, the bound structure of the protein was here considered as exactly known, while in a real-case docking, only an unbound structure of the protein can be known. Conformational changes between the bound and unbound protein are then likely to diminish the accuracy of the results.

Third, we considered the single-stranded state of the bound RNA as known a priori. In some cases, thisinformation is not available and must be retrieved from the modeling.

We consider several ways of overcoming these limits in our future research:

Regarding the scoring limitation, we have so far neglected: (i) the scoring specificities of fragments over full molecules, (ii) the internal energy of our fragments, and the compensation effect of e.g. breaking intra-fragment base stacking to permit stacking with protein residues, and (iii) the quality of pair connectivity, by using a simple boolean criterion. To increase the fraction of correct models in our initial sampling, we will (i) parametrize a new fragment-specific function to estimate the binding energy, (ii) sum of the internal and binding energies of each pose, (iii) weight the edges of the graph by the connectivity quality.

Regarding protein flexibility, we will use homology models with more or less divergence, MD or unbound forms (if available) in order to investigate its impact on the sampling quality of poses. If necessary, to model a flexible protein while avoiding the enumeration of its possible conformations, we will apply the same principle of decoupled sampling as for the RNA, but in a hierarchical way, by representing the protein as a tree of global and local conformations. Each "local" set of conformations can be used for docking, then the compatibility of conformations can be assessed together with the connectivity of the RNA poses bound to them.

To generalize the method to RNA of arbitrary base-pairing (secondary structure, "2D"), we will create a new 2D-specific fragment library, use a sequence-based prediction of a Boltzmann ensemble of 2D structures, dock all fragments of possible 2D structure at each sequence position, then incorporate the 2D probability in the assembly.

METHODS SUPPLEMENTARY

DeepATTRACT: As we dealt with a poly-U, one single docking of a UUU fragment was performed, and the poses were compared to each fragment at each position in the reference structure. POCASA was used with a 2 Å probe and a 1 Å spacing grid. Points with more than 500 neighbors within 7Å were retrieved. The fragment library was clustered with a 3 Å RMSD cutoff to keep representatives. Poses with an ATTRACT score below 100 kcal/Mol were kept for further testing with the whole corresponding cluster. The final poses were clustered with a 2 Å cutoff.

Partition function Z: For a pose i at position k, Z(k, i) is the sum, for all j connected to i, of {exp[(E(i) + E(j)) / RT] times Z(k-1, j)}, where E(i) is the ATTRACT score of pose i, and RT the Bolztman factor. In the stochastic backtracking, each pose is chosen with a probability Z(k,i) / sum(Z(k,j) over j).

Keywords: RNA modeling, RNA 3D structure, RNA, protein docking, combinatorial assembly